Universidad
Popular del Cesar

Minería de datos y Patrones

Version 2024-1

Selección de Características

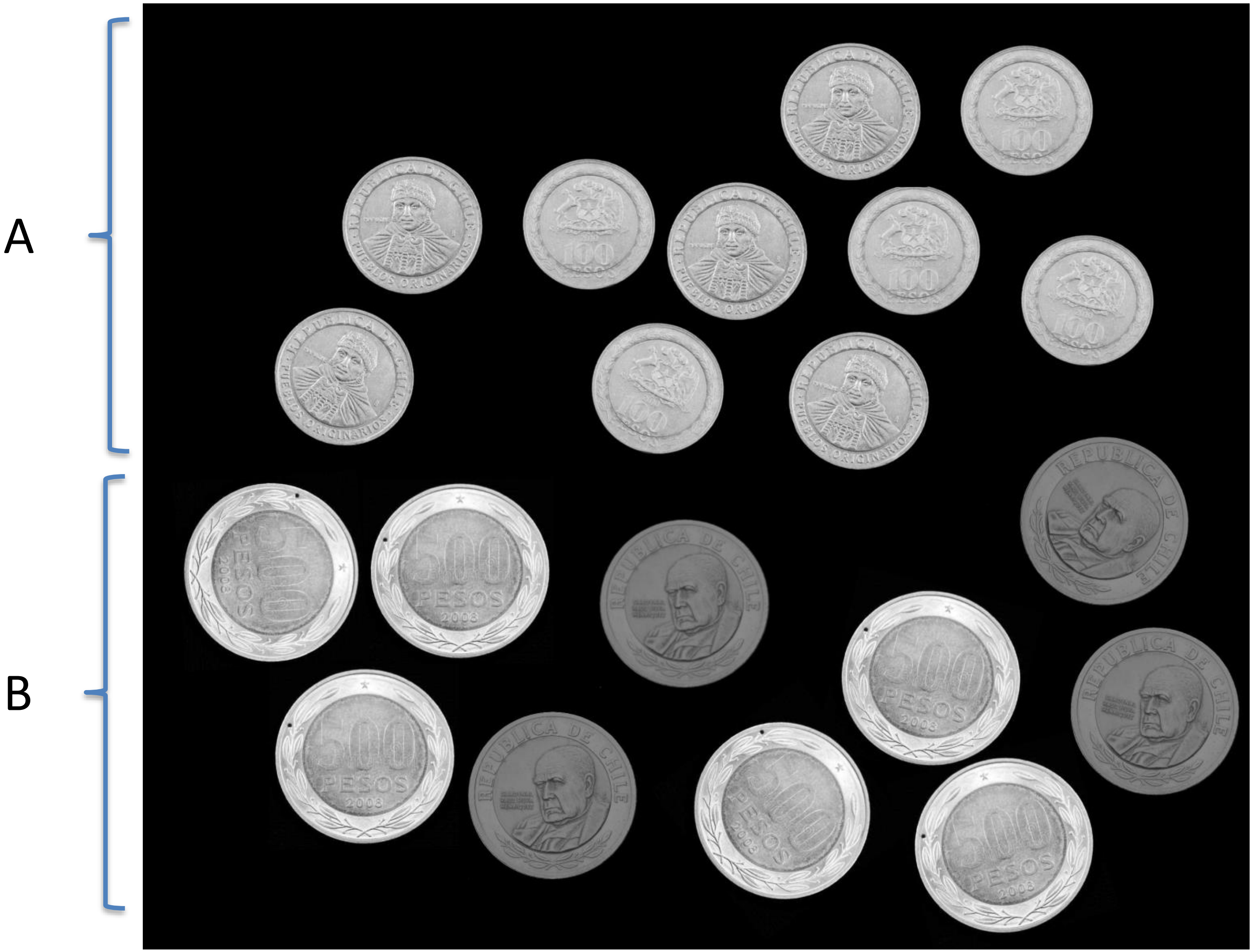
[Capítulo 3]

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

5 reasons why to select features

1. To avoid non-discriminative features
2. To avoid correlated features
3. To simplify the testing stage
4. To avoid false correlations
5. To avoid the curse of dimensionality

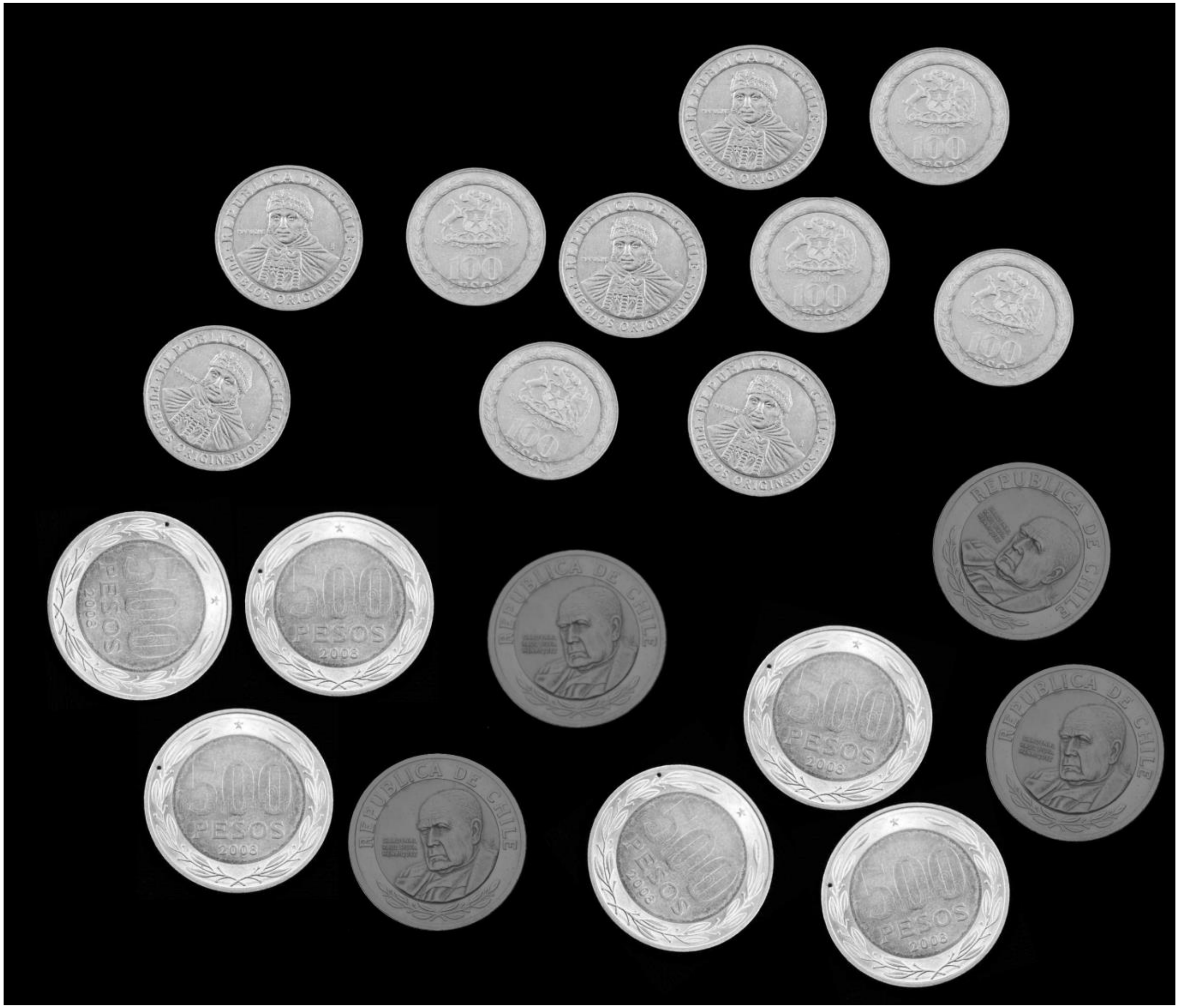
Example: Classification of Coins



5 reasons why to select features

1. To avoid non-discriminative features

1. To avoid non-discriminative features – ~~Grayvalue~~ average ☹



5 reasons why to select features

1. To avoid non-discriminative features
2. To avoid correlated features

2. To avoid correlated features – Area and ~~Diameter~~ ☹

5 reasons why to select features

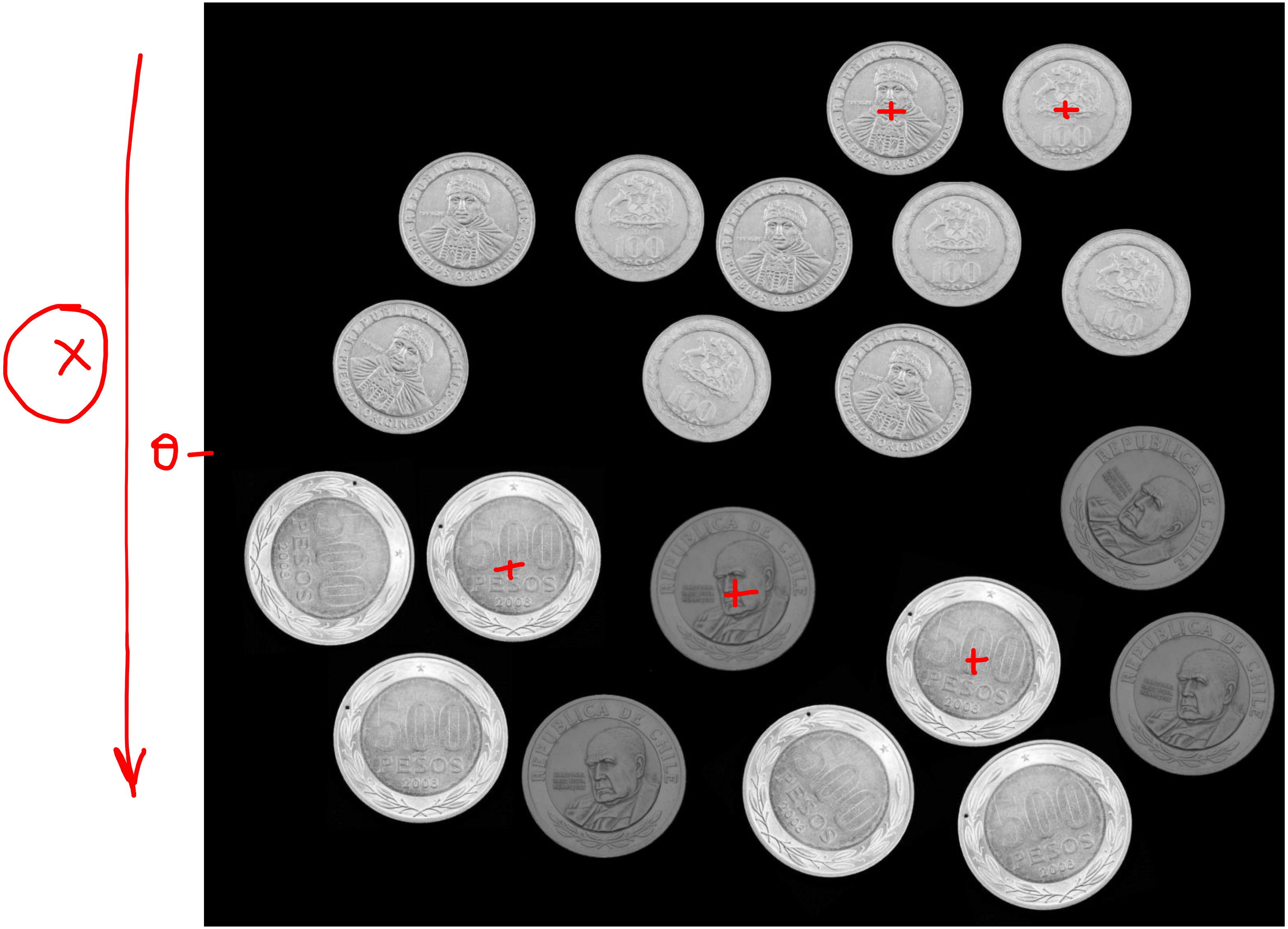
1. To avoid non-discriminative features
2. To avoid correlated features
3. To simplify the testing stage

3. To simplify the testing stage – Area vs. ~~Hu+Flusser~~

5 reasons why to select features

1. To avoid non-discriminative features
2. To avoid correlated features
3. To simplify the testing stage
4. To avoid false correlations

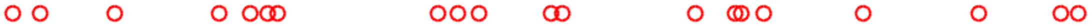
4. To avoid false correlations – ~~Location~~ ☹



5 reasons why to select features

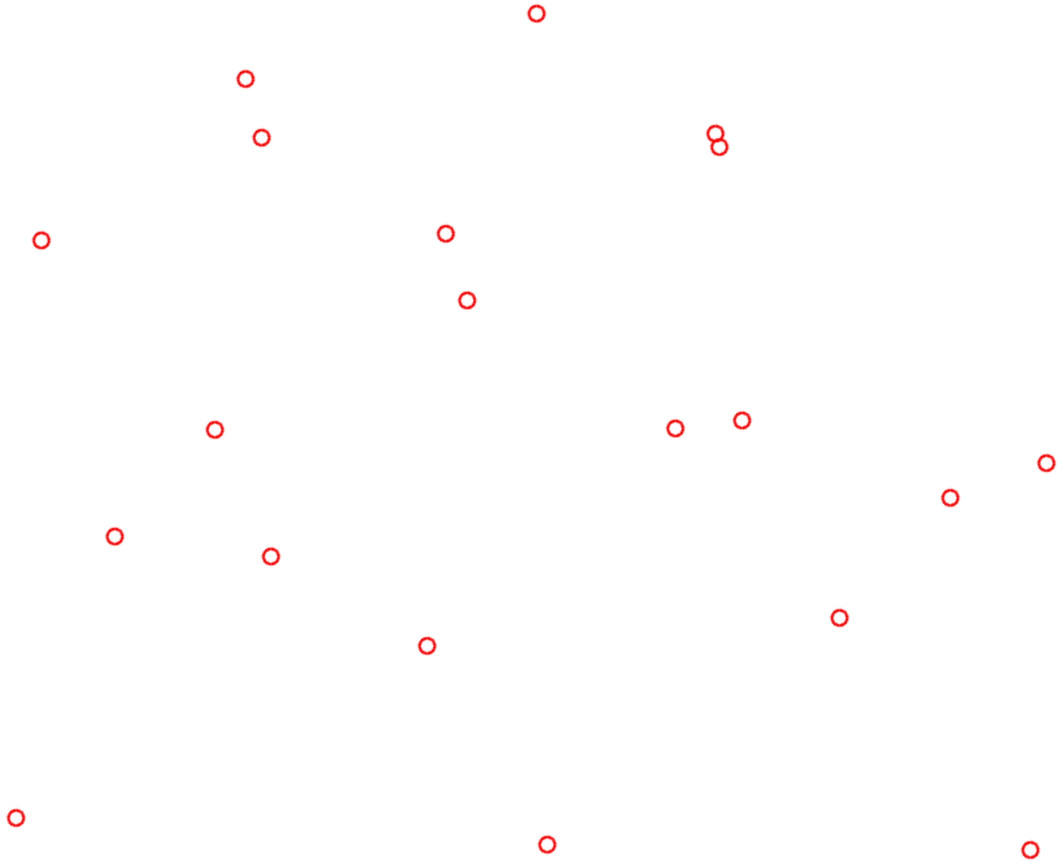
1. To avoid non-discriminative features
2. To avoid correlated features
3. To simplify the testing stage
4. To avoid false correlations
5. To avoid the curse of dimensionality

Curse of dimensionality – 1D



(20 points in 1D)

Curse of dimensionality – 2D

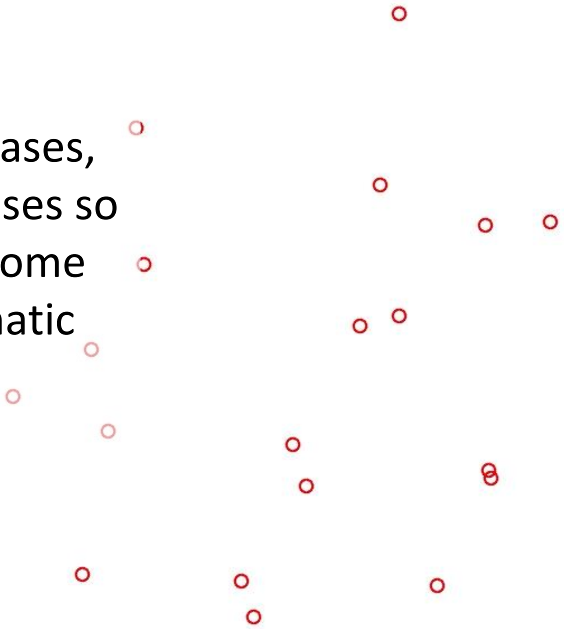


(20 points in 2D)

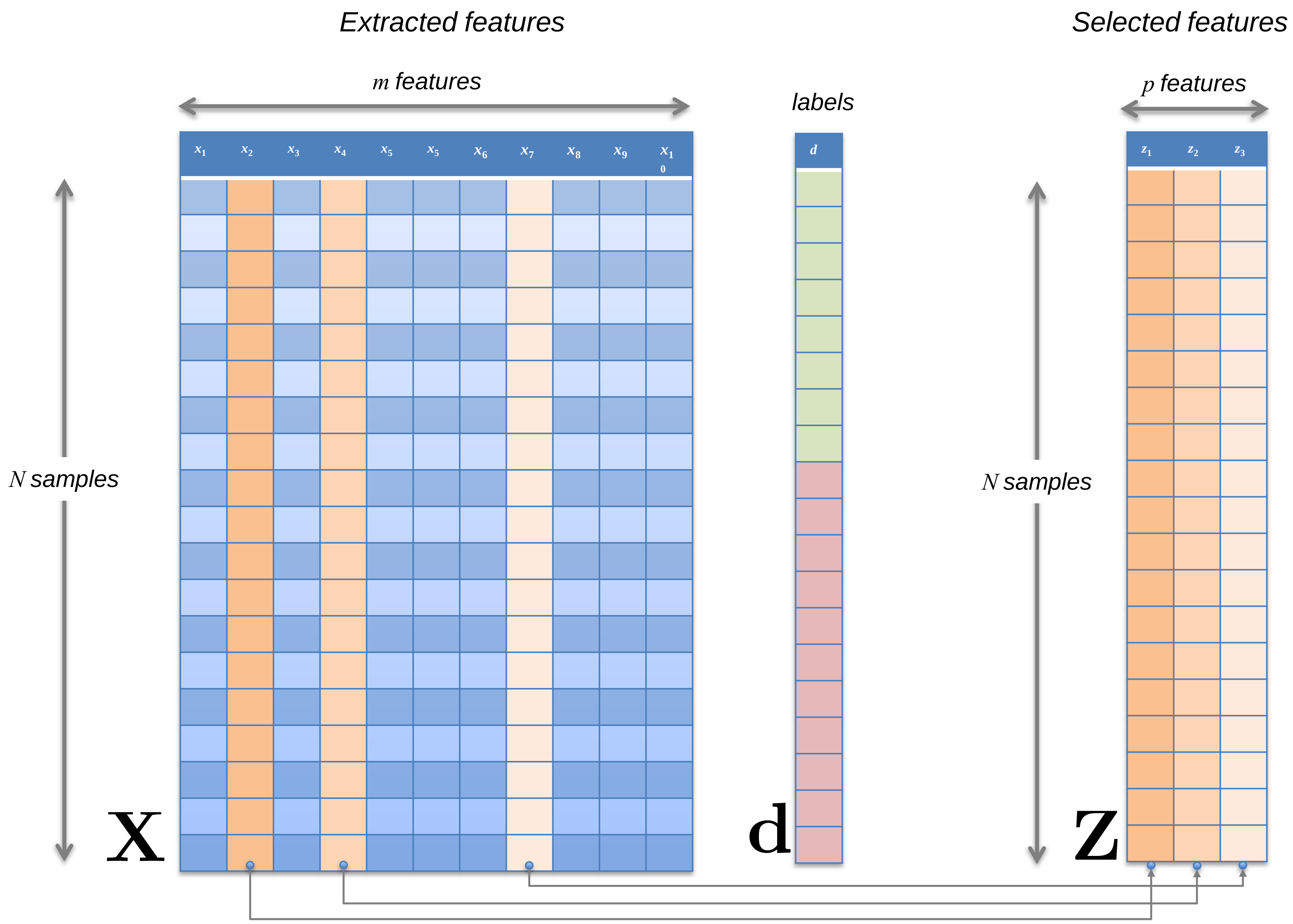
Curse of dimensionality – 3D

When the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance.

[Wikipedia]



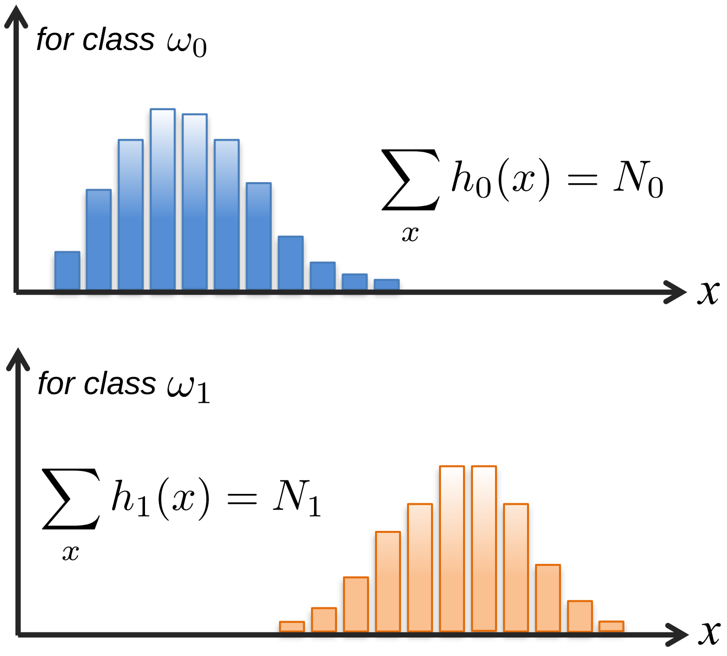
(20 points in 3D)



Extracted features

Selected features

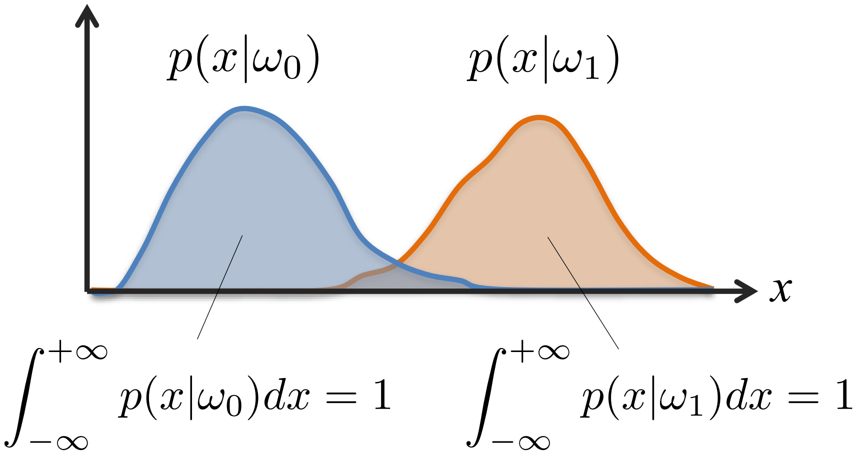
Frequency distributions



$$\sum_x h_0(x) = N_0$$

$$\sum_x h_1(x) = N_1$$

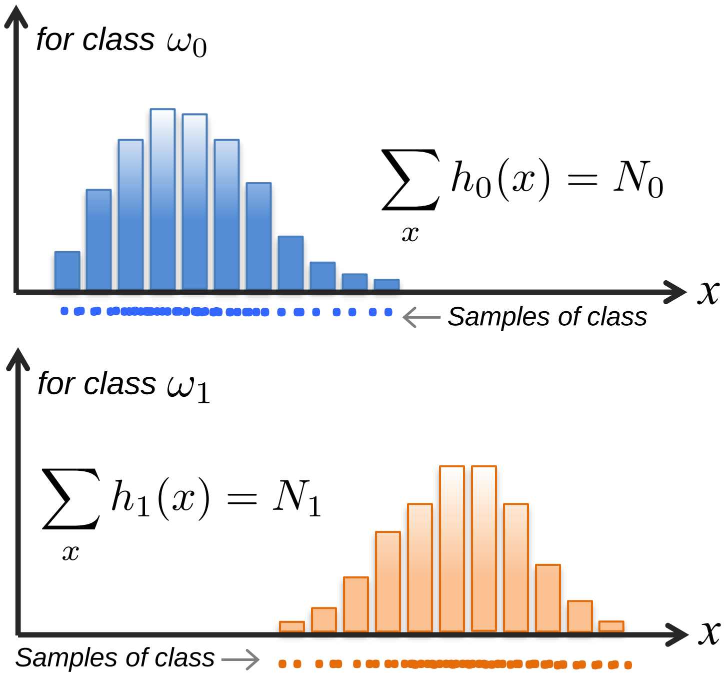
Probability density functions



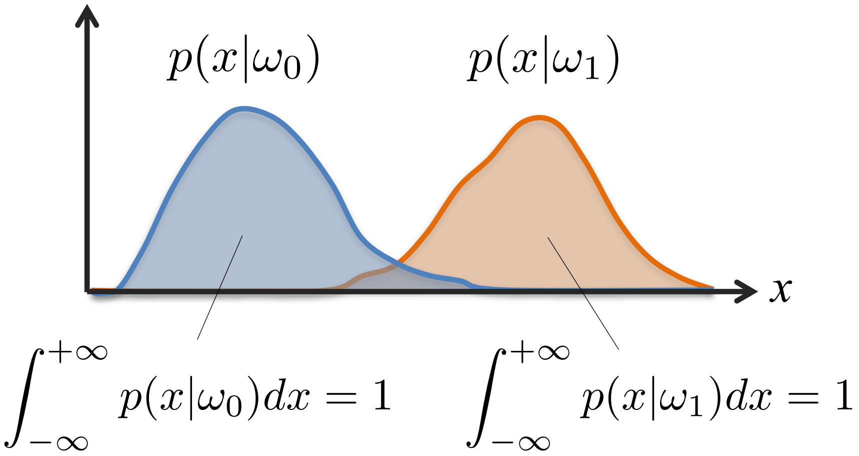
$$\int_{-\infty}^{+\infty} p(x|\omega_0)dx = 1$$

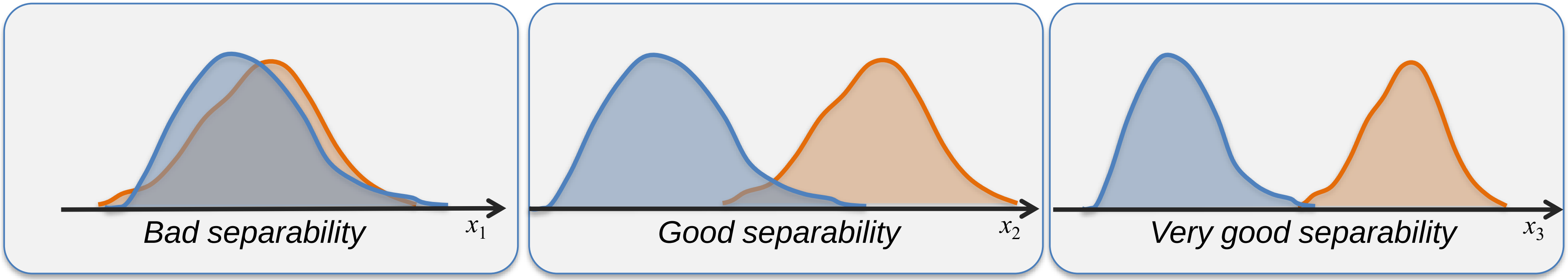
$$\int_{-\infty}^{+\infty} p(x|\omega_1)dx = 1$$

Frequency distributions

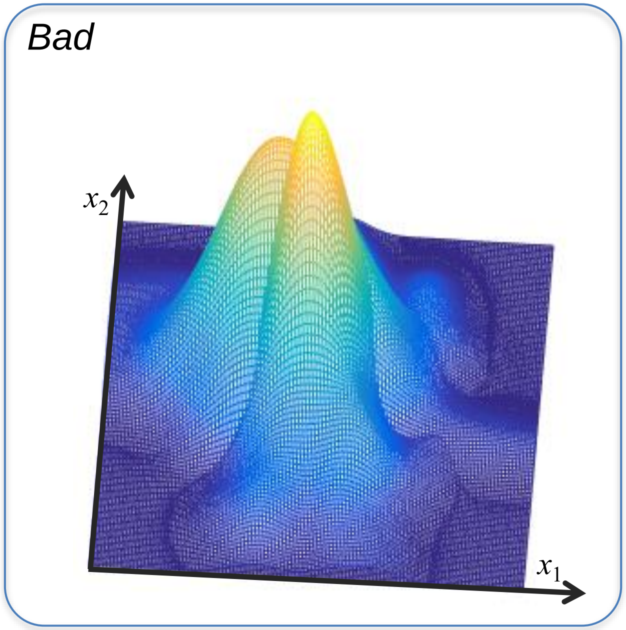


Probability density functions

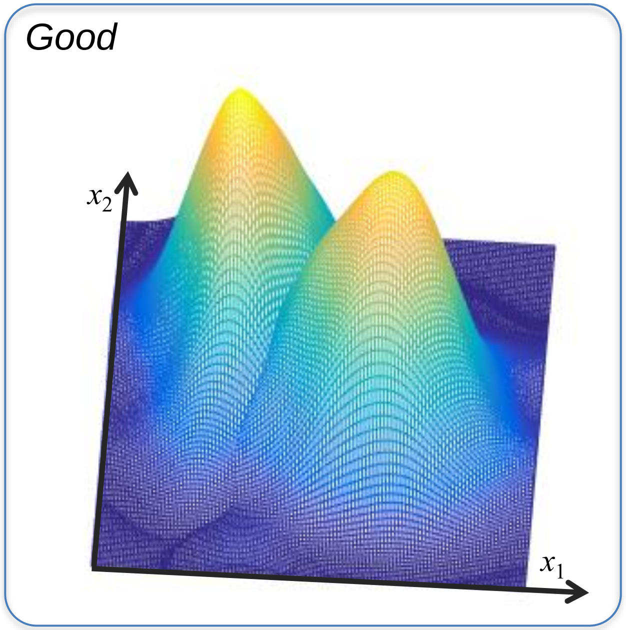




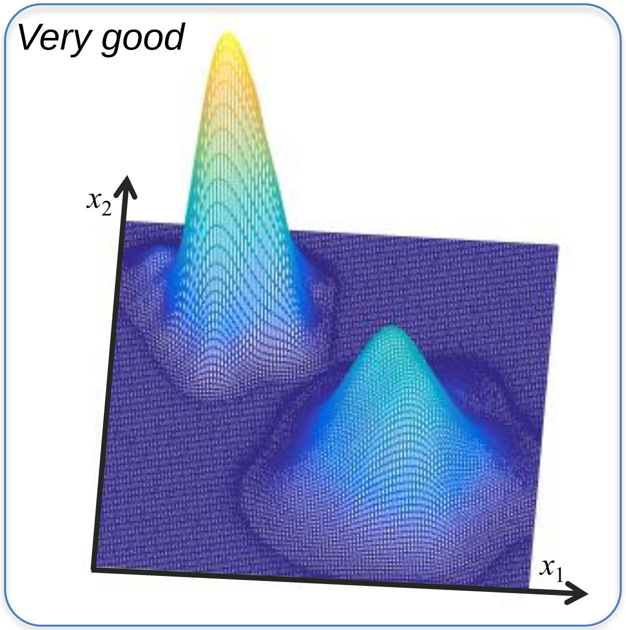
score = LOW　　score = MEDIUM　　score = HIGH



Bad

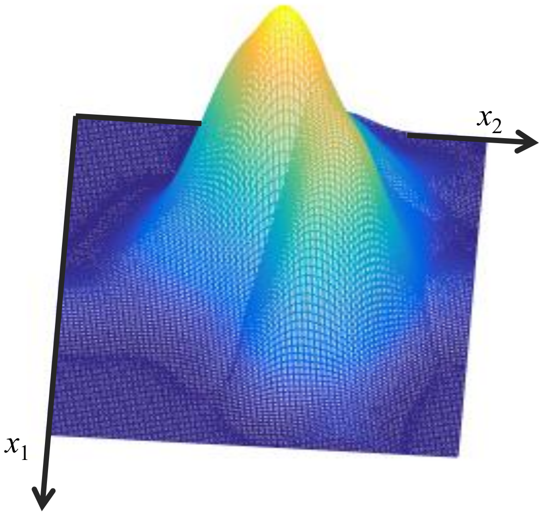


Good

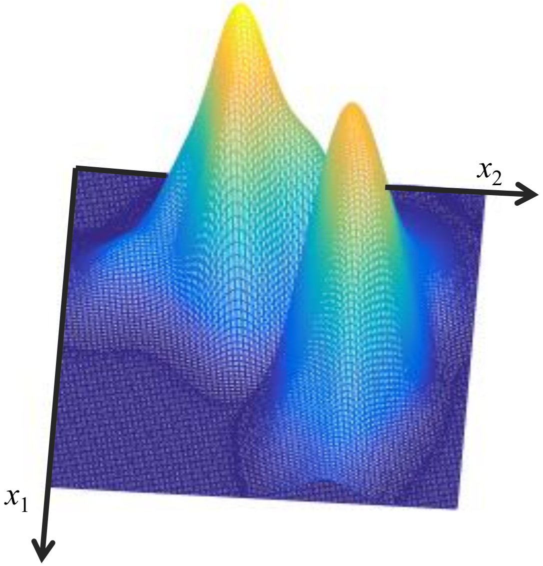


Very good

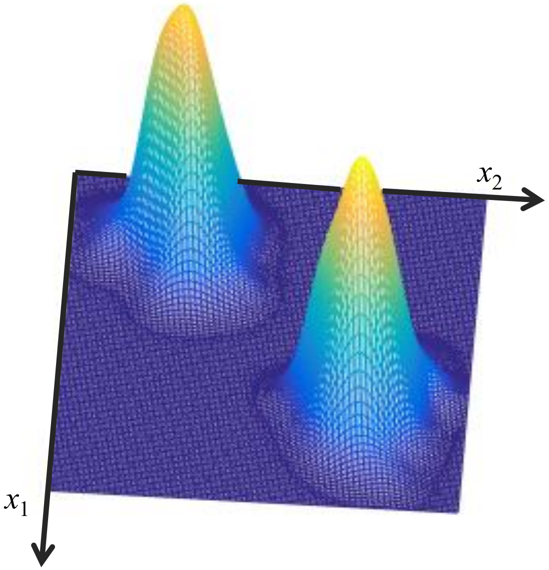
Bad

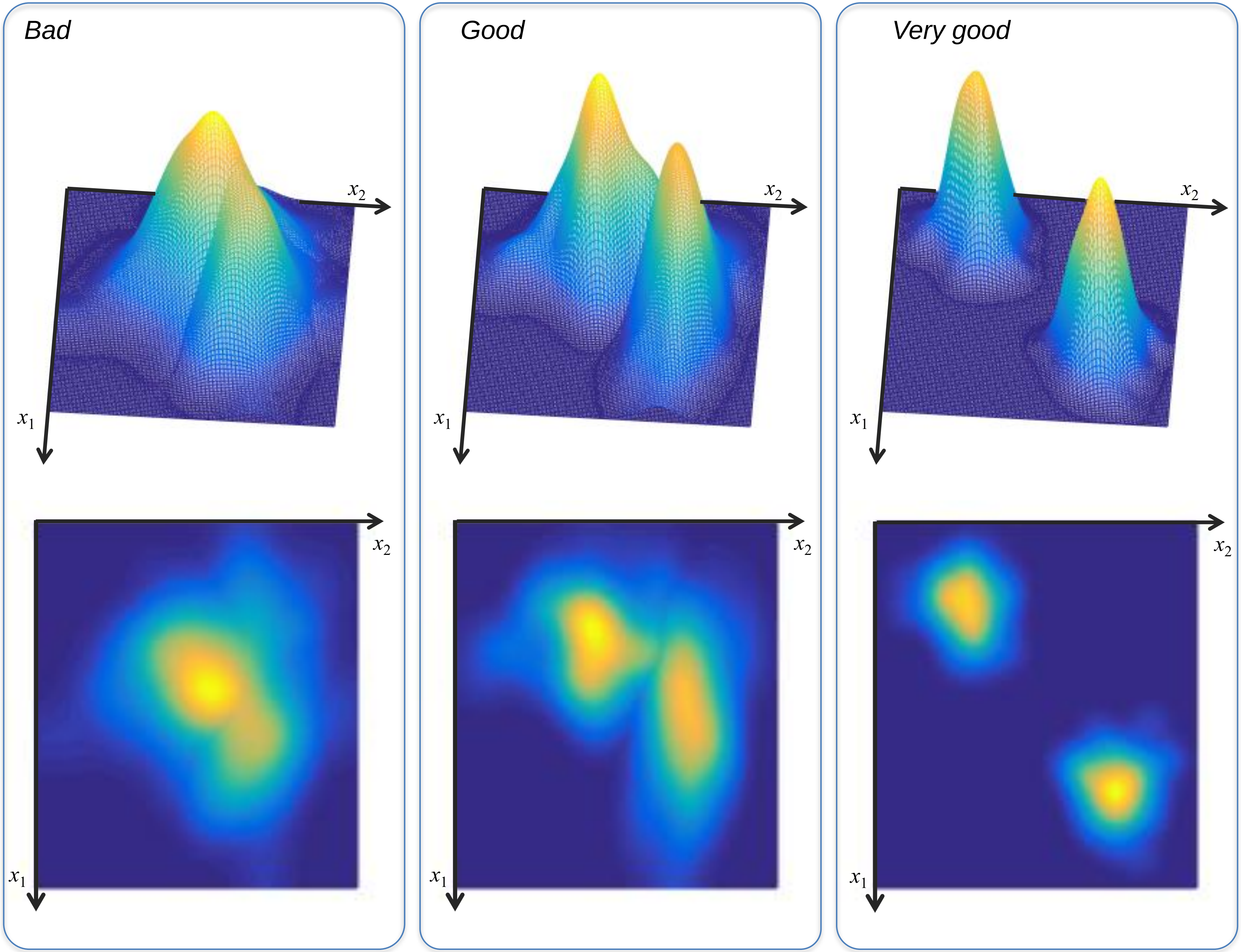


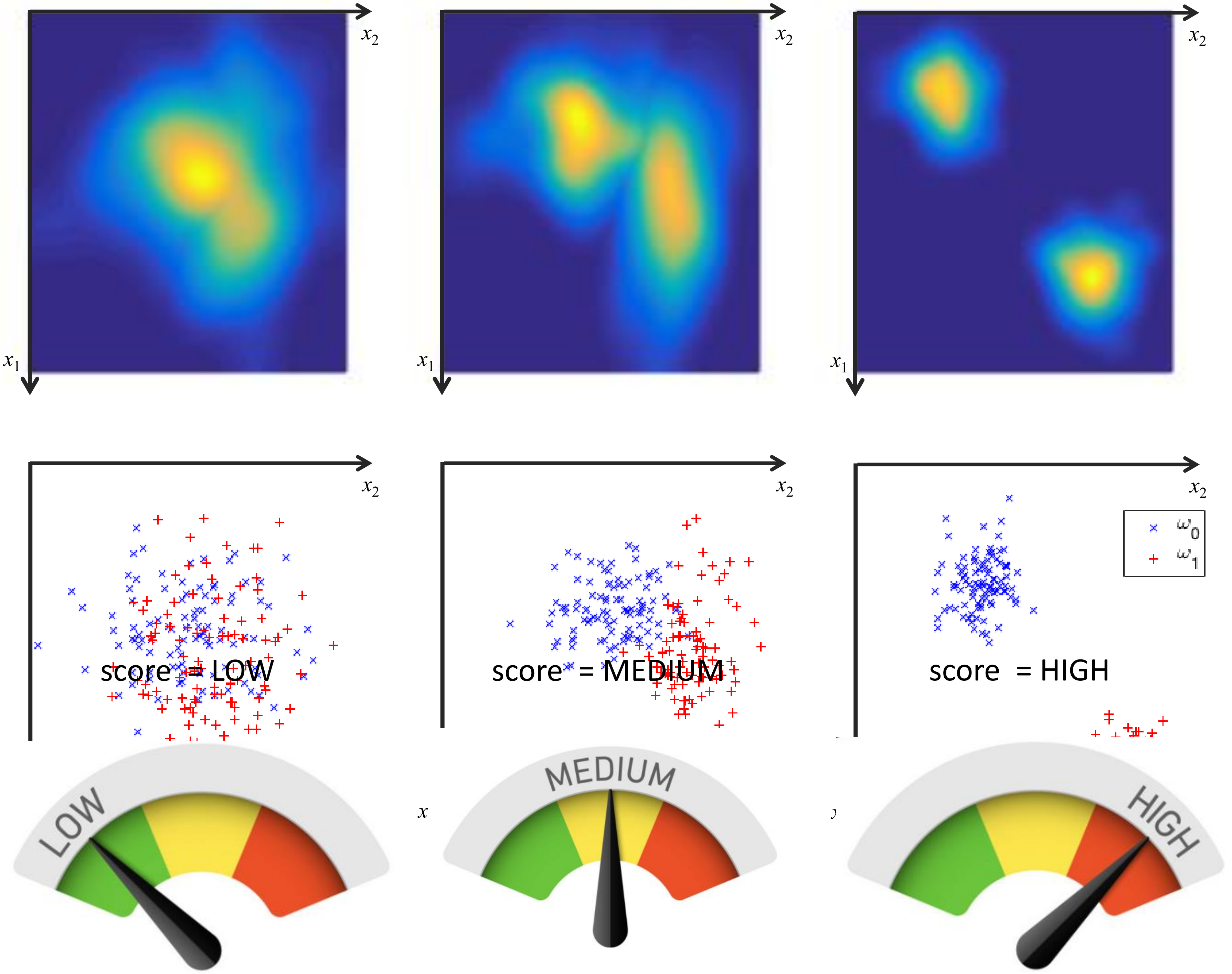
Good



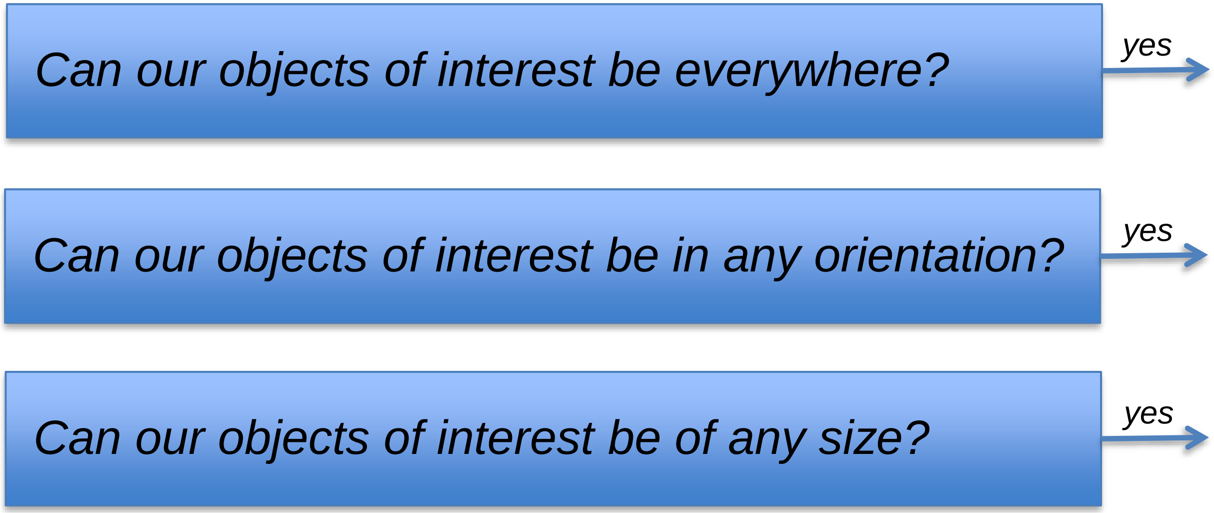
Very good




Bad
Good
Very good
x_2
x_1
x_2
x_1
x_2
x_1
x_2
x_1
x_2
x_1
x_2
x_1


x_2
x_1
x_2
x_1
x_2
x_1
x_2
x_2
x_2
ω_0
ω_1
score = LOW
score = MEDIUM
score = HIGH
LOW
MEDIUM
HIGH

Selected features must be…



Can our objects of interest be everywhere? *yes* → *invariant to translation*

Can our objects of interest be in any orientation? *yes* → *invariant to rotation*

Can our objects of interest be of any size? *yes* → *invariant to scale*